WIKIPEDIA

Principal component regression

In statistics, **principal component regression** (**PCR**) is a regression analysis technique that is based on principal component analysis (PCA). Typically, it considers regressing the outcome (also known as the response or the dependent variable) on a set of covariates (also known as predictors, or explanatory variables, or independent variables) based on a standard linear regression model, but uses PCA for estimating the unknown regression coefficients in the model.

In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors. One typically uses only a subset of all the principal components for regression, thus making PCR some kind of a regularized procedure. Often the principal components with higher variances (the ones based on eigenvectors corresponding to the higher eigenvalues of the sample variance-covariance matrix of the explanatory variables) are selected as regressors. However, for the purpose of predicting the outcome, the principal components with low variances may also be important, in some cases even more important.[1]

One major use of PCR lies in overcoming the multicollinearity problem which arises when two or more of the explanatory variables are close to being collinear.[2] PCR can aptly deal with such situations by excluding some of the low-variance principal components in the regression step. In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. This can be particularly useful in settings with high-dimensional covariates. Also, through appropriate selection of the principal components to be used for regression, PCR can lead to efficient prediction of the outcome based on the assumed model.

Contents

The principle

The PCR method may be broadly divided into three major steps:

1. Perform PCA on the observed data matrix for the explanatory variables to obtain the principal components, and then (usually) select a subset, based on some appropriate criteria, of the principal components so obtained for further use.
2. Now regress the observed vector of outcomes on the selected principal components as covariates, using ordinary least squares regression (linear regression) to get a vector of estimated regression coefficients (with dimension equal to the number of selected principal components).
3. Now transform this vector back to the scale of the actual covariates, using the selected PCA loadings (the eigenvectors corresponding to the selected principal components) to get the **final PCR estimator** (with dimension equal to the total number of covariates) for estimating the regression coefficients characterizing the original model.

Details of the method

Data Representation: Let $\mathbf{Y}_{n\times 1} = (y_1, \ldots, y_n)^T$ denote the vector of observed outcomes and $\mathbf{X}_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ denote the corresponding data matrix of observed covariates where, n and p denote the size of the observed sample and the number of covariates respectively, with $n \geq p$. Each of the n rows of $\mathbf{X}$ denotes one set of observations for the p dimensional covariate and the respective entry of $\mathbf{Y}$ denotes the corresponding observed outcome.

Data Pre-processing: Assume that $\mathbf{Y}$ and each of the p columns of $\mathbf{X}$ have already been centered so that all of them have zero empirical means. This centering step is crucial (at least for the columns of $\mathbf{X}$) since PCR involves the use of PCA on $\mathbf{X}$ and PCA is sensitive to centering of the data.

Underlying Model: Following centering, the standard Gauss–Markov linear regression model for $\mathbf{Y}$ on $\mathbf{X}$ can be represented as: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the unknown parameter vector of regression coefficients and $\boldsymbol{\varepsilon}$ denotes the vector of random errors with $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbf{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_{n\times n}$ for some unknown variance parameter $\sigma^2 > 0$

Objective: The primary goal is to obtain an efficient estimator $\widehat{\boldsymbol{\beta}}$ for the parameter $\boldsymbol{\beta}$, based on the data. One frequently used approach for this is ordinary least squares regression which, assuming $\mathbf{X}$ is full column rank, gives the unbiased estimator: $\widehat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ of $\boldsymbol{\beta}$. PCR is another technique that may be used for the same purpose of estimating $\boldsymbol{\beta}$.

PCA Step: PCR starts by performing a PCA on the centered data matrix $\mathbf{X}$. For this, let $\mathbf{X} = U\Delta V^T$ denote the singular value decomposition of $\mathbf{X}$ where, $\Delta_{p\times p} = \operatorname{diag}[\delta_1, \ldots, \delta_p]$ with $\delta_1 \geq \cdots \geq \delta_p \geq 0$ denoting the non-negative singular values of $\mathbf{X}$, while the columns of $U_{n\times p} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ and $V_{p\times p} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ are both orthonormal sets of vectors denoting the left and right singular vectors of $\mathbf{X}$ respectively.

The Principal Components: $V\Delta V^T$ gives a spectral decomposition of $\mathbf{X}^T\mathbf{X}$ where $\Lambda_{p\times p} = \operatorname{diag}[\lambda_1, \ldots, \lambda_p] = \operatorname{diag}[\delta_1^2, \ldots, \delta_p^2] = \Delta^2$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ denoting the non-negative eigenvalues (also known as the principal values) of $\mathbf{X}^T\mathbf{X}$, while the columns of V denote the corresponding orthonormal set of eigenvectors. Then, $\mathbf{X}\mathbf{v}_j$ and $\mathbf{v}_j$ respectively denote the j^{th} principal component and the j^{th} principal component direction (or PCA loading) corresponding to the j^{th} largest principal value λ_j for each $j \in \{1, \ldots, p\}$.

Derived covariates: For any $k \in \{1, \ldots, p\}$, let V_k denote the $p \times k$ matrix with orthonormal columns consisting of the first k columns of V. Let $W_k = \mathbf{X}V_k = [\mathbf{X}\mathbf{v}_1, \ldots, \mathbf{X}\mathbf{v}_k]$ denote the $n \times k$ matrix having the first k principal components as its columns. W may be viewed as the data matrix obtained by using the transformed covariates $\mathbf{x}_i^k = V_k^T\mathbf{x}_i \in \mathbb{R}^k$ instead of using the original covariates $\mathbf{x}_i \in \mathbb{R}^p \ \forall \ 1 \leq i \leq n$.

The PCR Estimator: Let $\widehat{\gamma}_k = (W_k^TW_k)^{-1}W_k^T\mathbf{Y} \in \mathbb{R}^k$ denote the vector of estimated regression coefficients obtained by ordinary least squares regression of the response vector $\mathbf{Y}$ on the data matrix W_k. Then, for any $k \in \{1, \ldots, p\}$, the final PCR estimator of $\boldsymbol{\beta}$ based on using the first k principal components is given by: $\widehat{\boldsymbol{\beta}}_k = V_k\widehat{\gamma}_k \in \mathbb{R}^p$.

Fundamental characteristics and applications of the PCR estimator

Two basic properties

The fitting process for obtaining the PCR estimator involves regressing the response vector on the derived data matrix W_k which has orthogonal columns for any $k \in \{1, \ldots, p\}$ since the principal components are mutually orthogonal to each other. Thus in the regression step, performing a multiple linear regression jointly on the k selected principal components as covariates is equivalent to carrying out k independent simple linear regressions (or univariate regressions) separately on each of the k selected principal components as a covariate.

When all the principal components are selected for regression so that $k = p$, then the PCR estimator is equivalent to the ordinary least squares estimator. Thus, $\widehat{\boldsymbol{\beta}}_p = \widehat{\boldsymbol{\beta}}_{\text{ols}}$. This is easily seen from the fact that $W_p = \mathbf{X}V_p = \mathbf{X}V$ and also observing that V is an orthogonal matrix.

Variance reduction

For any $k \in \{1, \ldots, p\}$, the variance of $\widehat{\boldsymbol{\beta}}_k$ is given by

$$\mathbf{Var}(\widehat{\boldsymbol{\beta}}_k) = \sigma^2 \, V_k(W_k^TW_k)^{-1}V_k^T = \sigma^2 \, V_k \, \operatorname{diag}\left(\lambda_1^{-1}, \ldots, \lambda_k^{-1}\right)V_k^T = \sigma^2 \sum_{j=1}^{k} \frac{\mathbf{v}_j\mathbf{v}_j^T}{\lambda_j}.$$

In particular:

$$\mathbf{Var}(\widehat{\boldsymbol{\beta}}_p) = \mathbf{Var}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) = \sigma^2 \sum_{j=1}^{p} \frac{\mathbf{v}_j\mathbf{v}_j^T}{\lambda_j}.$$

Hence for all $k \in \{1, \cdots, p-1\}$ we have:

$$\operatorname{Var}(\widehat{\beta}_{\text{ols}}) - \operatorname{Var}(\widehat{\beta}_k) = \sigma^2 \sum_{j=k+1}^{p} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}.$$

Thus, for ll $k \in \{1, \cdots, p\}$ we have:

$$\operatorname{Var}(\widehat{\beta}_{\text{ols}}) - \operatorname{Var}(\widehat{\beta}_k) \succeq 0$$

where $A \succeq 0$ indicates that a square symmetric matrix A is non-negative definite. Consequently, any given linear form of the PCR estimator has a lower variance compared to that of the same linear form of the ordinary least squares estimator.

Addressing multicollinearity

Under multicollinearity, two or more of the covariates are highly correlated, so that one can be linearly predicted from the others with a non-trivial degree of accuracy. Consequently, the columns of the data matrix $\mathbf{X}$ that correspond to the observations for these covariates tend to become linearly dependent and therefore, $\mathbf{X}$ tends to become rank deficient losing its full column rank structure. More quantitatively, one or more of the smaller eigenvalues of $\mathbf{X}^T\mathbf{X}$ get(s) very close or become(s) exactly equal to 0 under such situations. The variance expressions above indicate that these small eigenvalues have the maximum inflation effect on the variance of the least squares estimator, thereby destabilizing the estimator significantly when they are close to 0. This issue can be effectively addressed through using a PCR estimator obtained by excluding the principal components corresponding to these small eigenvalues.

Dimension reduction

PCR may also be used for performing dimension reduction. To see this, let L_k denote any $p \times k$ matrix having orthonormal columns, for any $k \in \{1, \ldots, p\}$. Suppose now that we want to approximate each of the covariate observations $\mathbf{x}_i$ through the rank k linear transformation $L_k \mathbf{z}_i$ for some $\mathbf{z}_i \in \mathbb{R}^k (1 \leq i \leq n)$.

Then, it can be shown that

$$\sum_{i=1}^{n} \|\mathbf{x}_i - L_k \mathbf{z}_i\|^2$$

is minimized at $L_k = V_k$, the matrix with the first k principal component directions as columns, and $\mathbf{z}_i = \mathbf{x}_i^k = V_k^T \mathbf{x}_i$, the corresponding k dimensional derived covariates. Thus the k dimensional principal components provide the best linear approximation of rank k to the observed data matrix $\mathbf{X}$.

The corresponding reconstruction error is given by:

$$\sum_{i=1}^{n} \left\|\mathbf{x}_i - V_k \mathbf{x}_i^k\right\|^2 = \begin{cases} \sum_{j=k+1}^{n} \lambda_j & 1 \leqslant k < p \\ 0 & k = p \end{cases}$$

Thus any potential dimension reduction may be achieved by choosing k, the number of principal components to be used, through appropriate thresholding on the cumulative sum of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Since the smaller eigenvalues do not contribute significantly to the cumulative sum, the corresponding principal components may be continued to be dropped as long as the desired threshold limit is not exceeded. The same criteria may also be used for addressing the multicollinearity issue whereby the principal components corresponding to the smaller eigenvalues may be ignored as long as the threshold limit is maintained.

Regularization effect

Since the PCR estimator typically uses only a subset of all the principal components for regression, it can be viewed as some sort of a regularized procedure. More specifically, for any $1 \leqslant k < p$, the PCR estimator $\widehat{\beta}_k$ denotes the regularized solution to the following constrained minimization problem:

$$\min_{\beta_* \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta_*\|^2 \quad \text{subject to} \quad \beta_* \perp \{\mathbf{v}_{k+1}, \cdots, \mathbf{v}_p\}.$$

The constraint may be equivalently written as:

$$V_{(p-k)}^T \beta_* = 0,$$

where:

$$V_{(p-k)} = [\mathbf{v}_{k+1}, \cdots, \mathbf{v}_p]_{p \times (p-k)}.$$

Thus, when only a proper subset of all the principal components are selected for regression, the PCR estimator so obtained is based on a hard form of regularization that constrains the resulting solution to the column space of the selected principal component directions, and consequently restricts it to be orthogonal to the excluded directions.

Optimality of PCR among a class of regularized estimators

Given the constrained minimization problem as defined above, let us consider the following generalized version of it:

$$\min_{\beta_* \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta_*\|^2 \quad \text{subject to} \quad L_{(p-k)}^T \beta_* = 0$$

where, $L_{(p-k)}$ denotes any full column rank matrix of order $p \times (p-k)$ with $1 \leqslant k < p$.

Let $\widehat{\beta}_L$ denote the corresponding solution. Thus

$$\widehat{\beta}_L = \arg\min_{\beta_* \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta_*\|^2 \quad \text{subject to} \quad L_{(p-k)}^T \beta_* = 0.$$

Then the optimal choice of the restriction matrix $L_{(p-k)}$ for which the corresponding estimator $\widehat{\beta}_L$ achieves the minimum prediction error is given by:[3]

$$L_{(p-k)}^* = V_{(p-k)} \Lambda_{(p-k)}^{1/2},$$

where

$$\Lambda_{(p-k)}^{1/2} = \operatorname{diag}\left(\lambda_{k+1}^{1/2}, \ldots, \lambda_p^{1/2}\right).$$

Quite clearly, the resulting optimal estimator $\widehat{\beta}_{L^*}$ is then simply given by the PCR estimator $\widehat{\beta}_k$ based on the first k principal components.

Efficiency

Since the ordinary least squares estimator is unbiased for β, we have

$$\operatorname{Var}(\widehat{\beta}_{\text{ols}}) = \operatorname{MSE}(\widehat{\beta}_{\text{ols}}),$$

where, MSE denotes the mean squared error. Now, if for some $k \in \{1, \cdots, p\}$, we additionally have: $V_{(p-k)}^T \beta = 0$, then the corresponding $\widehat{\beta}_k$ is also unbiased for β and therefore

$$\operatorname{Var}(\widehat{\beta}_k) = \operatorname{MSE}(\widehat{\beta}_k).$$

We have already seen that

$$\forall j \in \{1, \cdots, p\}: \quad \operatorname{Var}(\widehat{\beta}_{\text{ols}}) - \operatorname{Var}(\widehat{\beta}_j) \succeq 0,$$

which then implies:

$$\mathbf{MSE}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) - \mathbf{MSE}(\widehat{\boldsymbol{\beta}}_k) \succeq 0$$

for that particular k. Thus in that case, the corresponding $\widehat{\boldsymbol{\beta}}_k$ would be a more efficient estimator of $\boldsymbol{\beta}$ compared to $\widehat{\boldsymbol{\beta}}_{\text{ols}}$, based on using the mean squared error as the performance criteria. In addition, any given linear form of the corresponding $\widehat{\boldsymbol{\beta}}_k$ would also have a lower mean squared error compared to that of the same linear form of $\widehat{\boldsymbol{\beta}}_{\text{ols}}$.

Now Suppose now that for a given $k \in \{1, \cdots, p\}, V_{(p-k)}^T \boldsymbol{\beta} \neq \mathbf{0}$. Then the corresponding $\widehat{\boldsymbol{\beta}}_k$ is biased for $\boldsymbol{\beta}$. However, since

$$\forall k \in \{1, \cdots, p\} : \quad \mathbf{Var}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) - \mathbf{Var}(\widehat{\boldsymbol{\beta}}_k) \succeq 0,$$

it is still possible that $\mathbf{MSE}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) - \mathbf{MSE}(\widehat{\boldsymbol{\beta}}_k) \succeq 0$, especially if k is such that the excluded principal components correspond to the smaller eigenvalues, thereby resulting in lower bias.

In order to ensure efficient estimation and prediction performance of PCR as an estimator of $\boldsymbol{\beta}$, Park (1981) [3] proposes the following guideline for selecting the principal components to be used for regression: Drop the j^{th} principal component if and only if $\lambda_j < (p\sigma^2)/\boldsymbol{\beta}^T\boldsymbol{\beta}$. Practical implementation of this guideline of course requires estimates for the unknown model parameters σ^2 and $\boldsymbol{\beta}$. In general, they may be estimated using the unrestricted least squares estimates obtained from the original full model. Park (1981) however provides a slightly modified set of estimates that may be better suited for this purpose.[3]

Unlike the criteria based on the cumulative sum of the eigenvalues of $\mathbf{X}^T\mathbf{X}$, which is probably more suited for addressing the multicollinearity problem and for performing dimension reduction, the above criteria actually attempts to improve the prediction and estimation efficiency of the PCR estimator by involving both the outcome as well as the covariates in the process of selecting the principal components to be used in the regression step. Alternative approaches with similar goals include selection of the principal components based on cross-validation or the Mallows's Cp criteria. Often, the principal components are also selected based on their degree of association with the outcome.

Shrinkage effect of PCR

In general, PCR is essentially a shrinkage estimator that usually retains the high variance principal components (corresponding to the higher eigenvalues of $\mathbf{X}^T\mathbf{X}$) as covariates in the model and discards the remaining low variance components (corresponding to the lower eigenvalues of $\mathbf{X}^T\mathbf{X}$). Thus it exerts a discrete shrinkage effect on the low variance components nullifying their contribution completely in the original model. In contrast, the ridge regression estimator exerts a smooth shrinkage effect through the regularization parameter (or the tuning parameter) inherently involved in its construction. While it does not completely discard any of the components, it exerts a shrinkage effect over all of them in a continuous manner so that the extent of shrinkage is higher for the low variance components and lower for the high variance components. Frank and Friedman (1993)[4] conclude that for the purpose of prediction itself, the ridge estimator, owing to its smooth shrinkage effect, is perhaps a better choice compared to the PCR estimator having a discrete shrinkage effect.

In addition, the principal components are obtained from the eigen-decomposition of $\mathbf{X}$ that involves the observations for the explanatory variables only. Therefore, the resulting PCR estimator obtained from using these principal components as covariates need not necessarily have satisfactory predictive performance for the outcome. A somewhat similar estimator that tries to address this issue through its very construction is the partial least squares (PLS) estimator. Similar to PCR, PLS also uses derived covariates of lower dimensions. However unlike PCR, the derived covariates for PLS are obtained based on using both the outcome as well as the covariates. While PCR seeks the high variance directions in the space of the covariates, PLS seeks the directions in the covariate space that are most useful for the prediction of the outcome.

Recently, a variant of the classical PCR known as the **supervised PCR** was proposed by Bair, Hastie, Paul and Tibshirani (2006).[5] In a spirit similar to that of PLS, it attempts at obtaining derived covariates of lower dimensions based on a criteria that involves both the outcome as well as the covariates. The method starts by performing a set of p simple linear regressions (or univariate regressions) wherein the outcome vector is regressed separately on each of the p covariates taken one at a time. Then, for some $m \in \{1, \ldots, p\}$, the first m covariates that turn out to be the most correlated with the outcome (based on the degree of significance of the corresponding estimated regression coefficients) are selected for further use. A conventional PCR, as described earlier, is then performed, but now it is based on only the $n \times m$ data matrix corresponding to the observations for the selected covariates. The number of covariates used: $m \in \{1, \ldots, p\}$ and the subsequent number of principal components used: $k \in \{1, \ldots, m\}$ are usually selected by cross-validation.

Generalization to kernel settings

The classical PCR method as described above is based on classical PCA and considers a linear regression model for predicting the outcome based on the covariates. However, it can be easily generalized to a kernel machine setting whereby the regression function need not necessarily be linear in the covariates, but instead it can belong to the Reproducing Kernel Hilbert Space associated with any arbitrary (possibly non-linear), symmetric positive-definite kernel. The linear regression model turns out to be a special case of this setting when the kernel function is chosen to be the linear kernel.

In general, under the kernel machine setting, the vector of covariates is first mapped into a high-dimensional (potentially infinite-dimensional) feature space characterized by the kernel function chosen. The mapping so obtained is known as the feature map and each of its coordinates, also known as the feature elements, corresponds to one feature (may be linear or non-linear) of the covariates. The regression function is then assumed to be a linear combination of these feature elements. Thus, the underlying regression model in the kernel machine setting is essentially a linear regression model with the understanding that instead of the original set of covariates, the predictors are now given by the vector (potentially infinite-dimensional) of feature elements obtained by transforming the actual covariates using the feature map.

However, the kernel trick actually enables us to operate in the feature space without ever explicitly computing the feature map. It turns out that it is only sufficient to compute the pairwise inner products among the feature maps for the observed covariate vectors and these inner products are simply given by the values of the kernel function evaluated at the corresponding pairs of covariate vectors. The pairwise inner products so obtained may therefore be represented in the form of a $n \times n$ symmetric non-negative definite matrix also known as the kernel matrix.

PCR in the kernel machine setting can now be implemented by first appropriately centering this kernel matrix (K, say) with respect to the feature space and then performing a kernel PCA on the centered kernel matrix (K', say) whereby an eigendecomposition of K' is obtained. Kernel PCR then proceeds by (usually) selecting a subset of all the eigenvectors so obtained and then performing a standard linear regression of the outcome vector on these selected eigenvectors. The eigenvectors to be used for regression are usually selected using cross-validation. The estimated regression coefficients (having the same dimension as the number of selected eigenvectors) along with the corresponding selected eigenvectors are then used for predicting the outcome for a future observation. In machine learning, this technique is also known as *spectral regression*.

Clearly, kernel PCR has a discrete shrinkage effect on the eigenvectors of K', quite similar to the discrete shrinkage effect of classical PCR on the principal components, as discussed earlier. However, it should be noted that the feature map associated with the chosen kernel could potentially be infinite-dimensional, and hence the corresponding principal components and principal component directions could be infinite-dimensional as well. Therefore, these quantities are often practically intractable under the kernel machine setting. Kernel PCR essentially works around this problem by considering an equivalent dual formulation based on using the spectral decomposition of the associated kernel matrix. Under the linear regression model (which corresponds to choosing the kernel function as the linear kernel), this amounts to considering a spectral decomposition of the corresponding $n \times n$ kernel matrix $\mathbf{X}\mathbf{X}^T$ and then regressing the outcome vector on a selected subset of the eigenvectors of $\mathbf{X}\mathbf{X}^T$ so obtained. It can be easily shown that this is the same as regressing the outcome vector on the corresponding principal components (which are finite-dimensional in this case), as defined in the context of the classical PCR. Thus, for the linear kernel, the kernel PCR based on a dual formulation is exactly equivalent to the classical PCR based on a primal formulation. However, for arbitrary (and possibly non-linear) kernels, this primal formulation may become intractable owing to the infinite dimensionality of the associated feature map. Thus classical PCR becomes practically infeasible in that case, but kernel PCR based on the dual formulation still remains valid and computationally scalable.

See also

- Principal component analysis
- Partial least squares regression
- Ridge regression
- Multilinear subspace learning
- Canonical correlation
- Deming regression
- Total sum of squares

References

1. Jolliffe, Ian T. (1982). "A note on the Use of Principal Components in Regression". *Journal of the Royal Statistical Society, Series C*. **31** (3): 300–303. doi:10.2307/2348005 (https://doi.org/10.2307%2F2348005). JSTOR 2348005 (https://www.jstor.org/stable/2348005).
2. Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9
3. Sung H. Park (1981). "Collinearity and Optimal Restrictions on Regression Parameters for Estimating Responses". *Technometrics*. **23** (3): 289–295. doi:10.1080/00401706.1981.10487652 (https://doi.org/10.1080%2F00401706.1981.10487652).
4. Lldiko E. Frank & Jerome H. Friedman (1993). "A Statistical View of Some Chemometrics Regression Tools". *Technometrics*. **35** (2): 109–135. doi:10.1080/00401706.1993.10485033 (https://doi.org/10.1080%2F00401706.1993.10485033).
5. Eric Bair; Trevor Hastie; Debashis Paul; Robert Tibshirani (2006). "Prediction by Supervised Principal Components". *Journal of the American Statistical Association*. **101** (473): 119–137. doi:10.1198/016214505000000628 (https://doi.org/10.1198%2F016214505000000628).

Further reading

- Amemiya, Takeshi (1985). *Advanced Econometrics*. Harvard University Press. pp. 57–60. ISBN 0-674-00560-0.
- Theil, Henri (1971). *Principles of Econometrics*. Wiley. pp. 46–55. ISBN 0-471-85845-5.